IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 10, 2020, accepted July 8, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010248

The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation

TEWODROS LEGESSE MUNEA[1], YALEW ZELALEM JEMBRE[2], HALEFOM TEKLE WELDEGEBRIEL[1], LONGBIAO CHEN[1], CHENXI HUANG[1], AND CHENHUI YANG[1]

[1]School of Informatics, Xiamen University, Xiamen 361005, China
[2]Department of Electronic Engineering, Keimyung University, Daegu 42601, South Korea

Corresponding author: Chenhui Yang (chyang@xmu.edu.cn)

This work was supported in part by the "Double-First-Class Construction Fund".

ABSTRACT Human pose estimation localizes body keypoints to accurately recognizing the postures of individuals given an image. This step is a crucial prerequisite to multiple tasks of computer vision which include human action recognition, human tracking, human-computer interaction, gaming, sign languages, and video surveillance. Therefore, we present this survey article to fill the knowledge gap and shed light on the researches of 2D human pose estimation. A brief introduction is followed by classifying it as a single or multi-person pose estimation based on the number of people needed to be tracked. Then gradually the approaches used in human pose estimation are described before listing some applications and also flaws facing in pose estimation. Following that, a center of attention is given on briefly discussing researches with a significant effect on human pose estimation and examine the novelty, motivation, architecture, the procedures (working principles) of each model together with its practical application and drawbacks, datasets implemented, as well as the evaluation metrics used to evaluate the model. This review is presented as a baseline for newcomers and guides researchers to discover new models by observing the procedure and architecture flaws of existing researches.

INDEX TERMS Human pose estimation, pose estimation and action recognition, pose estimation survey, single and multi-person pose estimation.

I. INTRODUCTION

Human pose estimation is one of the challenging fields of study in computer vision which aims in determining the position or spatial location of body keypoints (parts/joints) of a person from a given image or video [1], [2], as shown in Fig.1. Thus, pose estimation obtains the pose of an articulated human body, which consists of joints and rigid parts using image-based observations [3].

Human pose estimation refers to the process of inferring poses in an image and these estimations are performed in either 3D or 2D [4]. To solve this problem, several approaches in the literature have been proposed. Early works introduced the classical approaches to articulated human pose estimation called the pictorial structures [5]–[8]. In these models, the spatial correlations of the body parts are demonstrated as a tree-structured graphical model and they are very successful when the limbs are visible however faced problems when the tree-structured fails capturing the correlation

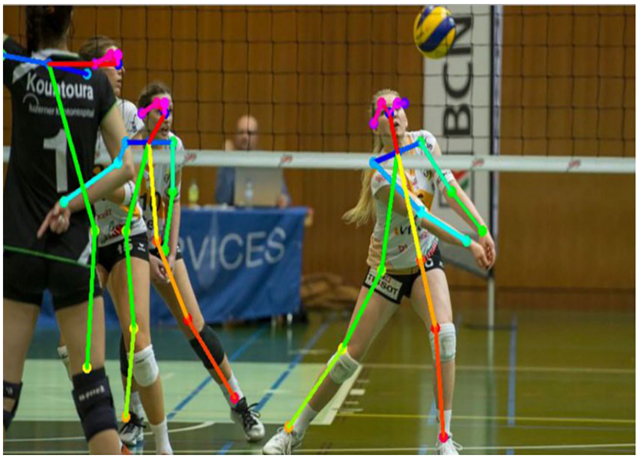

FIGURE 1. The estimated pose of each individual in a given image.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuping He.

between variables. Hand-crafted features such as edges, contours, the histogram of oriented gradients (HOG) features, and color histograms have also been used in early works for human pose estimation [9]–[13]. These models have shown bad generalization performance which faced problems in detecting the accurate location of the body parts.

Contributions: Solving the problems and challenges related to human pose estimation has been advanced and progressed remarkably with the help of deep learning and publicly available datasets. This survey provides a summary of these works comprehending up to date information and points the future research directions. Like some remarkable surveys [14]–[18], this paper also provides a general concept of human pose estimation. It can be used as a guideline for people who are new to this concept and helps them to define noble models by combining the network structures of the existing models. Additionally, it helps researchers to compare their work with significant models based on deep learning. Besides, here are some specific main contributions of this review:

- Provides a summary of preferred backbone architectures and loss functions used in addition to the overview of evaluation metrics implied and datasets employed.
- Provides an overview of recent models on 2D human pose estimation
- Limitations of each model's work and open issues are presented

A. RELATED WORKS

Other survey papers related to pose estimation have been released in the past years. For example, two survey papers [14], [15] published in 2016 have extensively surveyed models on human pose estimation which did not implement deep learning-based approaches. Then [16] presented a survey of deep learning, pose estimation, and application of deep learning for computer vision. A review on hand pose estimation is presented by [19] whereas [20] provided a survey on head pose estimation.

One of the recent surveys on 2D human pose estimation based on deep learning is [17]. This review started by categorizing pose estimation as a single person and multi-person pipeline and in each category created sub-categories. Another survey on deep-learning-based pose estimation has just come out [18] on both 2D and 3D pose estimation. 2D human pose estimation is categorized as [17] while 3D human pose estimation is categorized as model-free and model-based and the approaches are discussed based on these categories in both cases.

This survey paper presents different deep learning-based 2D human pose estimation models. The backbone architecture used, loss functions, the datasets used, as well as evaluation metrics implied are discussed and evaluated. The main objective of this paper is to provide a detailed analysis of mostly known effective models used, provide readers with various opportunities in mixing architecture of different models so that to come up with better human pose estimation models using better evaluation metrics or efficient backbone architecture.

B. BASIC STEPS IN POSE ESTIMATION

The main process of human pose estimation is boiled into two basic steps: i) localizing human body joints/keypoints and; ii) grouping those joints into valid human pose configuration [7], [8]. In the first step, the main focus is on finding the location of each keypoints of human beings as displayed in Fig.2. E.g. Head, shoulder, arm, hand, knee, ankle.

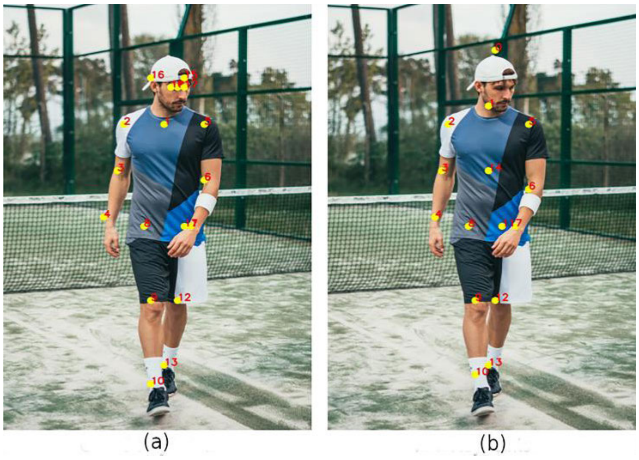

FIGURE 2. Keypoints localized by different dataset: (a) COCO keypoints, and (b) MPII keypoints.

Collecting and identifying these joints can be done through any of the different popular dataset formats; such that the way keypoints are stored in the selected dataset. As shown in Fig.2, different platforms can result in different dataset output formats for the same image of body joints. For instance, COCO [21], dataset provide 17 body joints whereas MPII [22] provides 14 body joints. Table 1 displays the outputs dataset for the two platforms.

The second step is grouping those joints into valid human pose configuration which determines the pairwise terms between body parts as seen in Fig.3. Different techniques have been applied in joining the keypoint candidates [23], [24].

The rest of this paper is organized as follows. Section II describes the category of pose estimation based on the number of people needed to track, approaches used in pose estimation, application of pose estimation, and flaws/challenges in pose estimation. Section III started by the introduction of backbone architectures used, the loss functions, dataset implied, and finally common evaluation metrics used to evaluate models. In section IV, a detailed discussion of each model's network procedures is discussed. Section V summarizes the models in short as a table and opens discussion based on presented in this article and finally section VI concludes the paper's works.

II. POSE ESTIMATION PRELIMINARY

This section discusses the general classification of pose estimation based on the number of people to track, introduce

TABLE 1. Different keypoints for MPII & COCO dataset.

COCO output format	**MPII** output format
Nose - 0	**Head** - 0
Neck - 1	Neck - 1
Right Shoulder - 2	Right Shoulder - 2
Right Elbow - 3	Right Elbow - 3
Right Wrist - 4	Right Wrist - 4
Left Shoulder - 5	Left Shoulder - 5
Left Elbow - 6	Left Elbow - 6
Left Wrist - 7	Left Wrist - 7
Right Hip - 8	Right Hip - 8
Right Knee - 9	Right Knee - 9
Right Ankle - 10	Right Ankle - 10
Left Hip - 11	Left Hip - 11
Left Knee - 12	Left Knee - 12
Left Ankle - 13	Left Ankle - 13
Right **Eye** - 14	**Chest** - 14
Left Eye - 15	
Right **Ear** - 16	
Left Ear - 17	
Background - 18	Background - 15

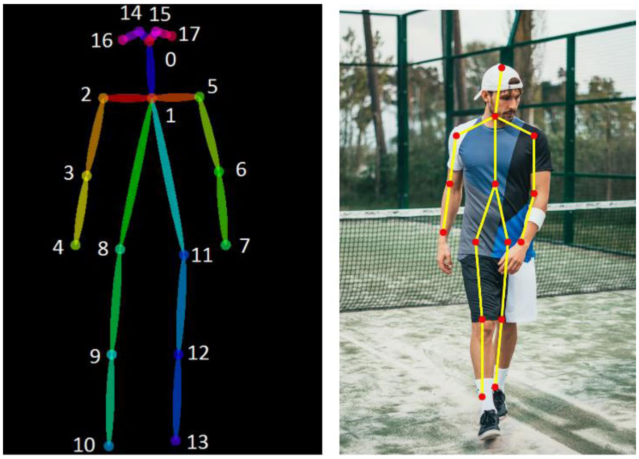

FIGURE 3. Configuration of valid human pose estimation.

the most popular approaches, application of pose estimation, and finally, the challenges that still require new as well as innovative approaches.

A. SINGLE/MULTI-PERSON POSE ESTIMATION

Based on the number of individuals being estimated given an image, pose estimation is classified as single-person and/or multi-person pose estimation. Single-person pose estimation is much easier compared to multi-person, to estimate pose for a single person from a given image which may contain (usually does) more than a single person. On the other hand, multi-person pose estimation determines the pose of all individuals available in the image [25]. Fig.4 shows the approach of a single person and multi-person pose estimation applied in the given images.

The technology of human pose estimation recently shown exciting progress on standard benchmarks both for a single person [26]–[30] and multi-person pose estimation [31]–[35].

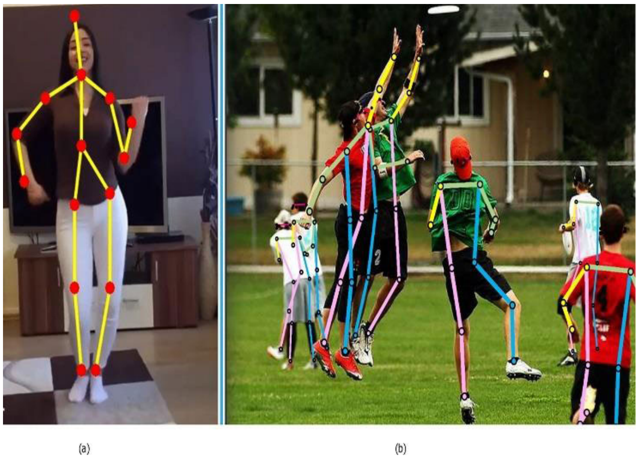

FIGURE 4. Classification of pose estimation: (a) Single person vs (b) multi-person pose estimation.

The use and introduction of deep learning-based architectures [36]–[39] and the availability of large-scale datasets such as MPII human pose dataset [22], COCO [21], and LSP [40] both single and multi-person pose estimation problems have lately been getting attention more and more.

B. APPROACHES IN POSE ESTIMATION

Two common approaches are employed in estimating the poses of individuals in a given image. 1) Top-down approaches, the processing is done from low to high resolutions, follow the detection of the individual instances in the image first using a bounding box object detector and then focus on determining their poses next [26], [27], [29], as shown in Fig.5.

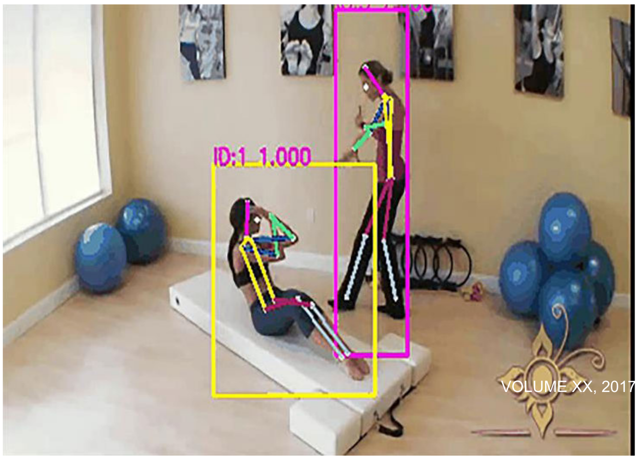

FIGURE 5. Pose estimation of multi-person in top-down approaches.

These approaches always suffer from early commitment, which means if the detection of individuals fails, there is no possibility of recovering. Also, it is vulnerable when multiple individuals are nearby. Furthermore, the computational cost depends on the number of people in the image, the more the people the more the computational cost. Hence, the run-time of these approaches is directly proportional to the number of people: means for every detection, a single-person pose estimator is run.

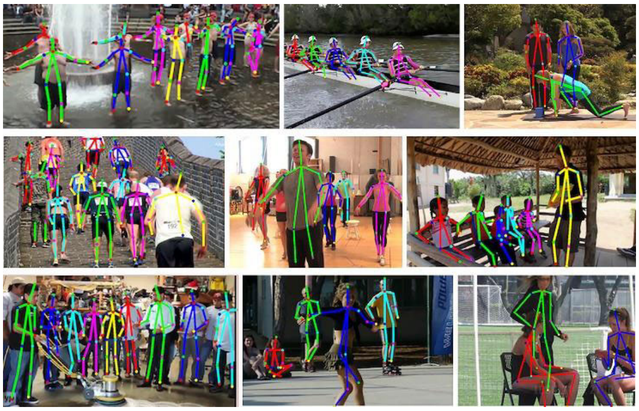

FIGURE 6. Some samples of multi-person pose estimation in bottom-up approaches.

2) The bottom-up approaches [31]–[34] processing is done from high to low resolutions. It starts by localizing identity-free semantic entities and then grouping them into person instance. Bottom-up approaches overcame early commitment and showed detached run-time complexity from the number of people in the image as shown in Fig.6. In addition to that, some researches using bottom-up approaches use the global contextual cues from other body parts and other people. However, bottom-up approaches face challenges in grouping body parts when there is a large overlap between people.

C. APPLICATIONS OF POSE ESTIMATION

Earlier human pose estimation application areas such as action recognition, human tracking, animation, and gaming [41], [42] are mentioned. Video surveillance, assisted living, and advanced driver assistance systems (ADAS) [43], [44] are also included. Furthermore, it may also provide game analysis in sports by describing the players' movement [45]. Pose estimation is also applicable in Sign languages to help disabled people. Some of the most common current applications of pose estimation are depicted in Fig.7.

D. CHALLENGES FACING

Principally every state-of-the-art (SOTA) pose estimation model includes a component that detects body joints or estimates their position and making pairwise terms between body part hypotheses which assist categorizing the pairwise terms into valid human pose configurations. In doing so, some challenges are faced. Such as position and scale of each person in the image; barely visible joints; interactions between people, which brings complex spatial interference due to clothing, lighting changes, contact, occlusion of individual parts by clothes, backgrounds, and limb articulations which makes the association of parts difficult. As the cost of 3D depth-sensing camera decreases and the machine learning algorithms to process the datasets of such technologies improve, we believe that it would bring new approaches to solve current challenges.

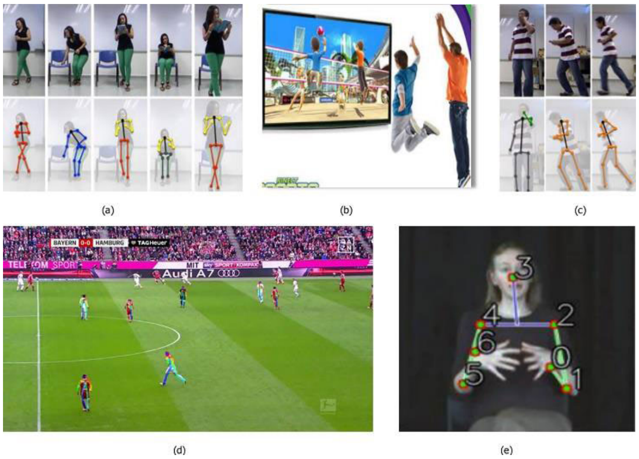

FIGURE 7. Applications of pose estimation: (a) action recognition, (b) gaming, (c) human tracking, (d) sports game analysis, and (e) sign languages.

III. MAIN COMPONENTS OF POSE ESTIMATION

Before diving into the details of each research model, better to explore first the main components of the pose estimation research's fundamentals such as backbone architecture, pose loss functions inhabited, the dataset used, and also evaluation metrics applied.

A. BACKBONE ARCHITECTURE

DeepPose [46] is the first significant research article that applied deep learning to human pose estimation. The authors have implemented the network architecture of AlexNet [37] as backbone architecture which consists of five convolution layers, two fully connected layers, and a softmax classifier. After the introduction of AlexNet, other machine learning algorithms such as R-CNN [47], Fast R-CNN [48], FPN [49], Faster R-CNN [50] and Mask R-CNN [39] have been used as backbone architecture for other human pose estimation researches [51], [32] and [52]. The second most popular backbone architecture is VGG [36] which has been used in [29], [34]. Although AlexNet and VGG have been in use for a while, most of the recent researches in human pose estimation [26], [27], [31], [32], [35], [53], have been using ResNet [38] as a backbone architecture.

B. LOSS FUNCTIONS

As one part of machine learning, human pose estimation models learn by loss functions. Loss functions evaluate how well a specific algorithm models the given dataset. It reduces the error in the prediction process [54], [55]. Largely three kinds of loss functions applied in human pose estimation models, namely, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Cross-Entropy loss.

MAE or L_1 loss function is calculated as the average of sums of all absolute differences between true and predicted values. L_1 loss function does not consider the direction, but only measures the magnitude of the error. The L_1 loss function is not sensitive to outliers thus it is robust. But it is very hard to regress precisely which brings complexity for machine learning [56].

$$L_1 = 1/n \sum_{i=1}^{n} |y_i - f(x_i)| \quad (1)$$

MSE also called L_2 loss or Quadratic loss function is calculated as the average of the squared differences between true and predicted values. Like L_1 loss, the L_2 loss function measures the magnitude of error without considering the direction. L_2 loss function provides an easier way to calculate gradients due to its mathematical properties. But, the L_2 loss function is very sensitive to outliers, unlike L_1 because of its usage of squaring when predicted values and true values are very far away occasionally [56].

$$L_2 = 1/n \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad (2)$$

In Cross-Entropy loss (Negative Log-Likelihood or Log loss), each predicted probability is compared to the actual class output value (0 or 1) and a score is calculated that penalizes the probability based on the distance from the expected value [54]. The penalty is logarithmic, offering a small score for small differences (0.1 or 0.2) and an enormous score for a large difference (0.9 or 1.0) [54]. This means an algorithm with smaller Cross-Entropy loss is preferable, and if it has 0.0 Log loss, then it predicts perfect probability.

$$Log_{loss} = -(y_i log(f(x_i)) + (1 - y_i) log(1 - f(x_i))) \quad (3)$$

C. DATASET

Researchers in human pose estimation have been mainly using the following four datasets which are freely available to the public: FLIC, LSP, MPII Human Pose, and COCO. Less known datasets such as Pascal VOC [57], SURREAL [58] for single-person in both 2D and 3D pose estimation, HumanEva, Human3.6 dataset, CrowdPose [59], and JTA [60] have also been used in human pose estimation.

Frames Labeled In Cinema (FLIC) [61] dataset consists of a total of 5,003 images of which 80% (around 4,000 images) are used as training and 20% (around 1016 images) are used as testing dataset. FLIC dataset is acquired from popular 30 movies in Hollywood by running a person detector SOTA model on every tenth frame of 30 movies. These images contain individuals in different kinds of poses with different kinds of clothing. From the dataset, each individual is labeled with 10 body joints. In most cases, the FLIC dataset has been used for a single person and multi-person pose estimation models.

Leeds Sports Pose dataset (LSP) [40] and LSP Ext (LSP extension or sometimes expressed as LSPe) contain a combination of 11,000 training and 1,000 testing images from Flickr. These images are mostly from sports activities which make it very challenging in their appearance terms. In addition to that, most individuals in the image have scaled to roughly 150 pixels in length. In the LSP dataset, each individual's full body is labeled with a total of 14 joints which shows an increased number of joints compared to FLIC. To be specific, the LSP dataset has a total of 2,000 annotated images whereas LSP Ext has a total of 10,000 images. In most cases, both datasets have been used for single person pose estimation models.

Max Planck Institute for Informatics (MPII) Human Pose dataset [22] contains around 25,000 images from which composed of more than 40,000 individuals with annotated body joints. These images are collected on the purpose to show human activities every day. In MPII human pose dataset, each individual's body is labeled with 15 body joints as mentioned in the introduction section. As FLIC dataset, MPII Human pose dataset has also been used for a single person and multi-person pose estimation models.

Finally, the MS-COCO dataset has got huge attention for multi-person pose estimation models. MS-COCO or usually called COCO (Common Objects in Context) is a product of Microsoft (MS) [21]. COCO dataset is a collection of a very large dataset with annotation types of object detection, keypoint detection, stuff segmentation, panoptic segmentation, and image captioning. A JSON file is used to store annotations. COCO dataset brought to the table a very interesting mix of data, with various human poses used in different body scales, also containing occlusion patterns, with unconstrained environments. COCO dataset contains a total of 200,000 images and these contain 250,000 people with keypoints from which each individual's instance is labeled with 17 joints. COCO dataset has been producing dataset starting from 2014 with a large amount.

D. COMMON EVALUATION METRICS

Similar to any other research, human pose estimation also uses evaluation metrics to compare and contrast one model from the other. Some researchers claim the superiority of their model based on a metric they developed, which could lead to false performance improvement. This section glances some of the commonly used evaluation metrics in human pose estimation research, such that a consistent result is presented in the field and also researchers new to the field can easily adapt to these metrics.

1) Percentage of Correct Parts (PCP): this metric measures the detection rate of limbs. A limb (body part) is considered detected if the distance between the two predicted joint locations and the true limb joint locations is less than half of the limb length [46]. PCP commonly referred also as PCP@0.5. Recently, PCP has not been preferred as an evaluation metric even though it was initially regarded as the go-to metric. This is because PCP penalizes shorter limbs. The higher the PCP the better the model.

2) Percentage of Detected Joints (PDJ): this metric is proposed to address the limitations observed in PCP. This evaluation metric defines a joint correctly detected if the distance between the predicted joint location and the true joint location is within a certain fraction of the torso diameter (the distance between the right hip and left shoulder). For instance, for PDJ@0.2, it means the distance between the pre-

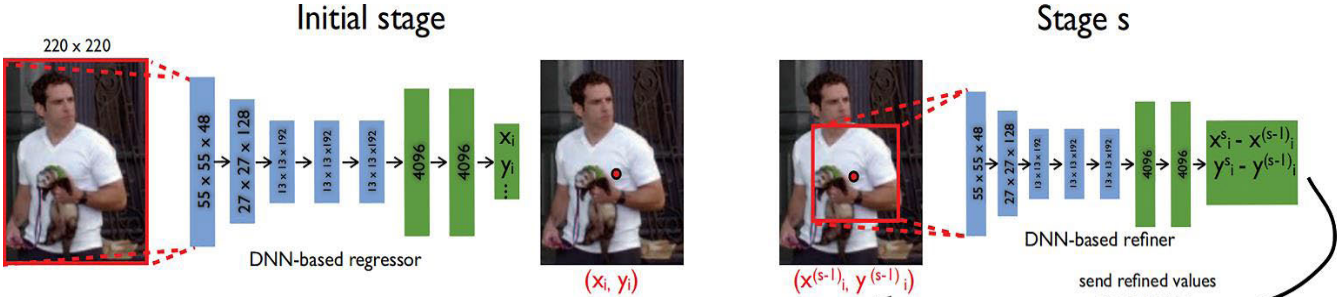

FIGURE 8. DeepPose's DNN-based pose regressor and refiner.

dicted joint location and the true joint location should be less than 0.2 * torso diameter. By changing this fraction, detection rates are obtained for different degrees of localization precision.

3) Percentage of Correct Key-points (PCK): this also measures the distance between the predicted joint location and the true joint location. The PCK evaluation metric measures the body joints' localization accuracy. The criteria of PCK and PDJ are very similar except to the fact that the torso diameter is replaced with the maximum side length (or threshold) of the external rectangle of ground truth body joints [62]. Thus, detecting a joint is considered correct if the distance between the predicted joint and the true joint is within a certain fraction of the specified threshold. Again, the higher the PCK the better the model.

4) PCKh is a modified version of PCK. PCKh's matching threshold is 50% of the head segment length (a portion of the head length is used as a reference at 50%). PCKh is also defined as the head-normalized probability of the correct keypoint metric [63]. In PCKh, joint detection is considered correct if the predicted joint location is with a certain threshold from the true joint location. But the threshold should be adaptively selected based on the individual's size. It should fall within αl pixels of the ground-truth position, where α is a constant and l is the head size that corresponds to 60% of the diagonal length of the ground-truth head bounding box. The PCKh@0.5 ($\alpha = 0.5$) score is reported [63]. To make the metric articulation independent, one probably better chooses to use the head size.

5) Area Under the Curve (AUC) measures the different range PCK thresholds (E.g., when α varies from 0 to 0.5) entirely. It informs how the model is capable of distinguishing each body's joints. The higher the AUC, the better the model.

6) Object Keypoint Similarity (OKS) gives a measure of how a predicted keypoint is close to ground truth. OKS is much similar to IoU (Intersection over Union) in keypoint detection performance. So, when a model gets higher OKS, it means the overlap between the predicted keypoint and the truth is higher.

Besides the above evaluation metrics, Average Precision (AP) and mean Average Precision (mAP) are also used [64], [65]. In addition to deep learning being advanced and having very large datasets, non-linear jumping systems discussed in [66]–[68] can also help to improve the efficiency of different algorithms in deep learning models.

IV. MAJOR RESEARCHES IN HUMAN POSE ESTIMATION

We will now dive into some unique and most effective pose estimation models' network flow. While discussing each approach we will explain, how their architecture is organized? How CNN architecture got deeper [69]? Is it a single/multi-person model? what kind of loss functions the models are using? dataset implied and how evaluated the work?

A. DeepPose

DeepPose [46], a single person pose estimation model published in 2014, formulated body joints as a problem of a CNN-based regression (which is a class of DNN-based regression). The authors have used AlexNet [37] as a backbone architecture, to analyze the effects of jointly training a multi-staged architecture with repeated intermediate supervision. DeepPose refines the coarse pose to get better estimation using a cascade of regressors which output coordinates (x, y) of each joint. When joints are predicted in DeepPose cascaded regressors, images are cropped around that joint to feed for the next stage. This allows the subsequent regressors to learn features for finer scales because a higher resolution images guide them to better precision.

DeepPose has a total of 3-stages cascade of regressors to estimate the pose of an individual in a given image. The Overall network structure of DeepPose is shown in Fig.8, in which the blue color shows the convolutional layers while the green shows the fully connected layers. The left schematic view shows the initial stage which contains the DNN-based regressor for the coarse pose. When joints are predicted at this stage, the image is cropped around the coordinates of the detected joint then passed to the next stage called the DNN-based refiner (on the right side of Fig.8) as an input.

The performance of this model is evaluated on two datasets (FLIC and LSP) using evaluation metrics of PCK and PCP. This model outperformed the previous SOTA works in most cases.

Even though producing the first CNN based human pose estimation model is very significant, the work has some limitations. The main problem was regressing to a location is very difficult. This increased the complexity of the learning

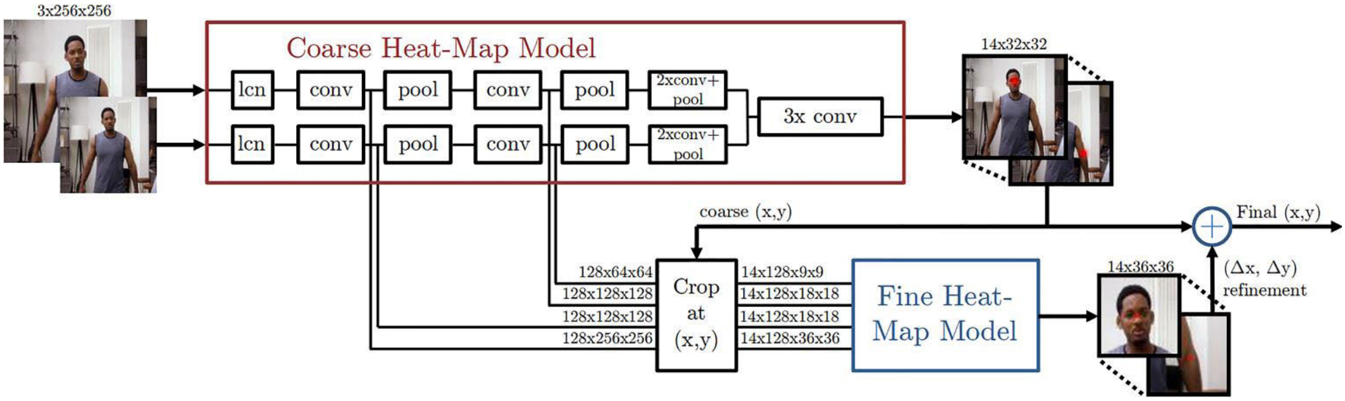

FIGURE 9. ConvNet pose overall network structure.

which weakened generalization. Thus, DeepPose performed very poorly in some regions. However, it has been very helpful for recent SOTA researches to change the challenge to the problem of estimating heatmaps for available joints or keypoints.

B. ConvNet POSE: EFFICIENT OBJECT LOCALIZATION USING CONVOLUTIONAL NETWORKS

In this paper [70], ConvNet architecture, multi-resolution CNN architecture is proposed to generate discrete heatmaps instead of continuous regression that predicts the probability of the location of individual joints in monocular RGB images. In ConvNet pose, different scale features are captured simultaneously using multiple resolution CNN architectures in parallel.

This model implements a sliding window detector that produces a coarse heatmap output and this coarse heatmap is refined by 'pose refinement' ConvNet to get better localization which improves in recovering the spatial accuracy lost due to pooling in the initial model. This means the model contains a module (a convolutional network) for coarse localization, a module for sampling and cropping the features of ConvNet for each joint at a specified location (x, y), and also a module for fine-tuning as shown in Fig.9 which displays the model's overall network structure.

This model has shown the use of a ConvNet and a graphical model jointly. The spatial relationship between the joints is typically learned by the graphical model [71]. The performance of the model is evaluated using PCK and PCKh@0.5 on FLIC [61] and MPII [22] dataset respectively in which outperformed the previous SOTA models.

This model implemented the joint use of a convolutional network and graphical model. Also, it revealed heatmaps are preferable than direct joint regression. Human poses are structured because of physical connections (like knees are rigidly related to hips and ankles), body part proportions, joint limits (like knees do not bend forward), left-right symmetries, interpenetration constraints, and others. Thus, modeling this structure realizes that detecting visible keypoints is easier and this directs on estimating the occluded keypoints which are very hard to detect. However, this model lacks structure modeling.

C. CPM: CONVOLUTIONAL POSE MACHINES

CPM [29] consists of a sequence of convolutional networks that produce a 2D belief map for the location of each keypoint. The sequential prediction framework provided by CPM helps them to learn rich implicit spatial information and feature representation of images at the same time. CPM is completely differentiable and the multi-stage architecture can be trained end-to-end. Thus, the image features and belief maps produced by the previous stage are given as input for the next stage (except the first stage) in CPM. One of the basic motivations for CPM is learning long-range spatial relationships and this is done using large receptive fields. Also, CPM used intermediate supervision after each stage to avoid the problem of vanishing gradients.

The overall network structure and receptive field of CPM are shown in Fig.10. CPM network is divided into multiple stages (the stage is used as hyper-parameter, usually =3) and at each stage, the confidence (belief) map of each keypoint is computed. Fig.10, (a) and (b) show the structures in the pose machine, (c) and (d) show the corresponding convolutional networks respectively, while (e) shows the receptive fields at different stages.

At the first stage a basic convolutional network, a classic VGG structure represented by X predicts the belief maps of each keypoint from the original input image. This leads to the condition that if the individual in the image has p joint points, then the belief map has p layers with each layer representing the joint point heatmap. Each layer's loss is added up as a total loss to achieve intermediate supervision which helped them in vanishing gradients.

For subsequent stages, stage ≥ 2, the structure is the same except the input to the network are two data: a belief map output from the previous stage and the result of the original image passed through X'. In addition to that, CPM showed that increasing the receptive field increases the accuracy of the prediction of keypoints. Furthermore, CPM implemented

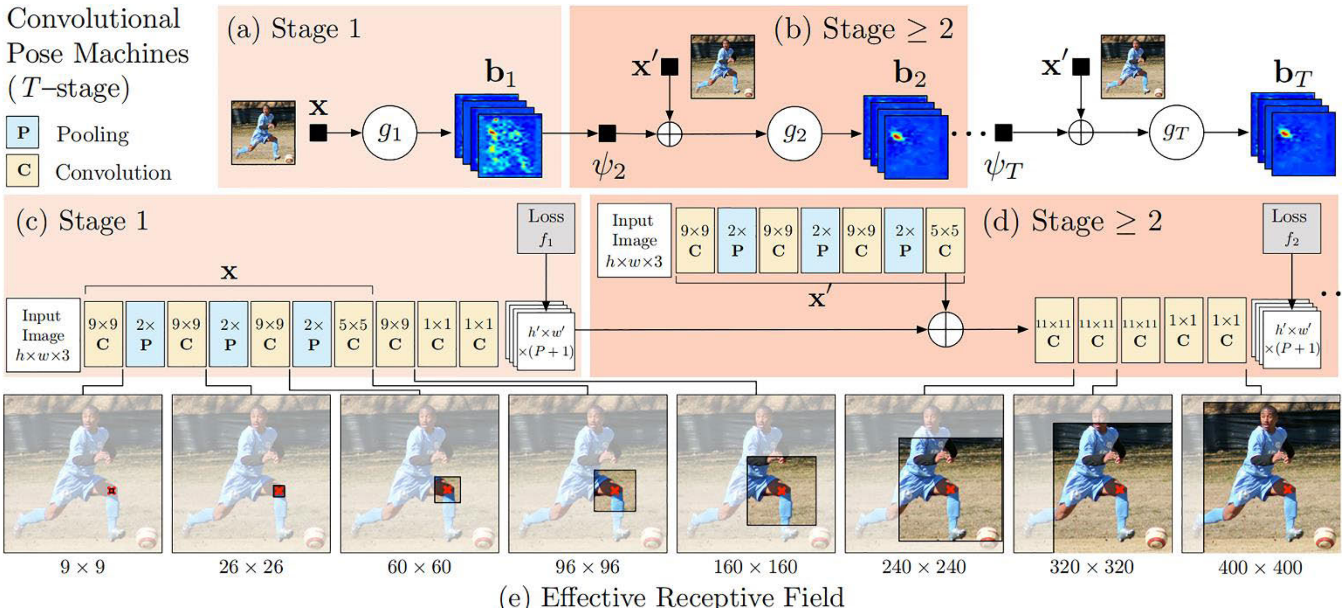

FIGURE 10. Network architecture and receptive field of CPM.

intermediate supervision after each stage to solve the vanishing gradients.

CPM implemented their model on three known datasets: MPII, LSP, and FLIC using evaluation metrics of PCK@0.1, PCK@0.2, and PCKh@0.5. It is noteworthy to mention that CPM achieved a PCKh@0.5 score of 10.76% higher than the previous SOTA on the most challenging part, the ankle.

CPM is the integration of the convolutional network to pose machines to learn image features and image-dependant spatial models to estimate human poses. Nevertheless, this work implemented a top-down approach on single person pose estimation, which leads to known errors and complexities of the top-down approach discussed earlier.

D. STACKED HOURGLASS NETWORKS FOR HUMAN POSE ESTIMATION

Stacked hourglass network [27], exactly lookalike of an hourglass stacked which are composed as steps of pooling and upsampling layers, is on the basic motivation of capturing information at every scale. In human pose estimation: an individual's orientation, limb arrangements, the relationship between adjacent joints, and other many cues that are best identified at different scales in a given image.

Thus, the stacked hourglass network is performing a repeated use of bottom-up (from high resolution to low resolution using pooling), top-down (from low resolution to high resolution using upsampling), and intermediate supervision to improve the network performance. The overall network structure of stacked hourglass modules is given in Fig.12. The hourglass stacked helps them to capture information on every scale means both global and local information is captured. It means skip connections are used to preserve spatial information in every resolution and pass it for upsampling.

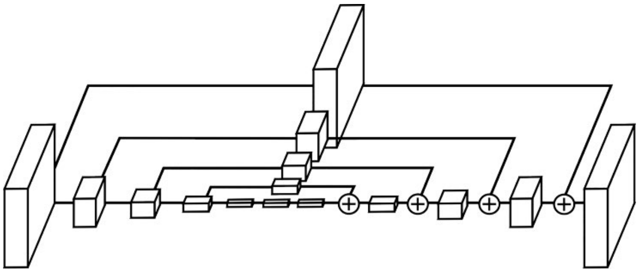

FIGURE 11. A single hourglass module in which a box represents a residual module.

Fig.11 displays a single hourglass module in which a single box represents a residual module. The primary module in an hourglass structure, residual or recurrent learning, is used for the bypass addition structure. This residual learning, composed of three convolutional layers with different scales in which batch normalization and ReLu inserted between them, extracts higher-level features while maintaining the primary level of information. The second path skips the path and contains only one kernel, A convolution layer with a scale of 1. Thus, only the data depth is changed not the data size.

For each hourglass module, a fourth-order residual module is used. The 4th-order Hourglass sub-network extracts features from the original scale to the 1/16 scale. It does not change the data size, only the data depth. The hourglass module is used to capture local information contained in pictures at different scales. At different scales, it may contain a lot of useful information, such as the position of the human body, the movements of the limbs, the relationship between adjacent joint points, and so on. First, the Conv layer and Max Pooling layer are used to scale features to a small resolution. At each Max Pooling (down-sampling), the network forks (branches) and convolves the features with

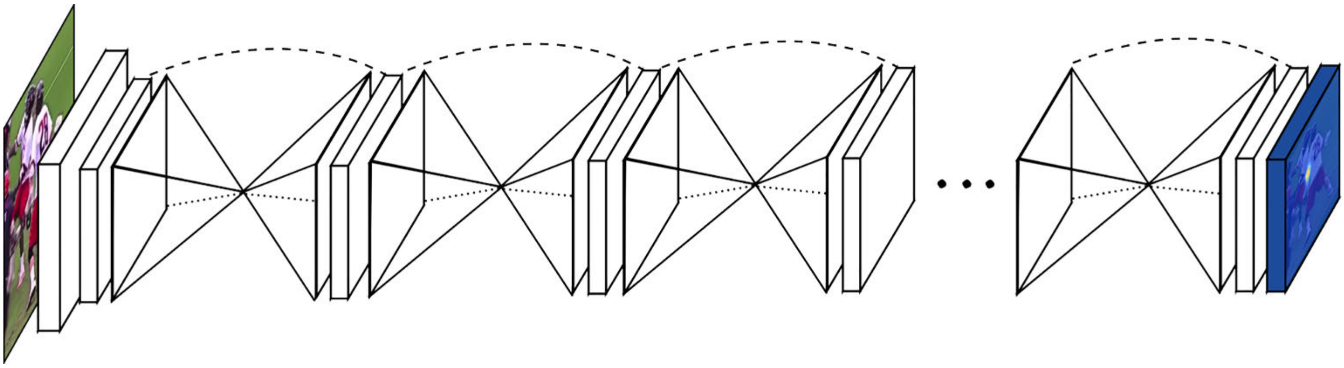

FIGURE 12. The overall structure of stacked hourglass modules.

the original pre-pooled resolution; After getting the lowest resolution features, the network starts up-sampling, and gradually combines feature information of different scales. The lower resolution here uses the nearest neighbor upsampling method, and two different feature sets are added element by element (which performs two different feature sets Add elements).

In stacked hourglass down-sampling uses max pooling, and up-sampling uses nearest-neighbor interpolation. The original image is down-sampled and input into the Hourglass sub-net. The output of Hourglass goes through two linear modules to get the final response graph. During this period, the Residual module and the convolutional layer are used to gradually extract features. The secondary used network centered around two Hourglass and repeats the second (latter) half of the primary network. The input of the second Hourglass contains three channels (paths): the input data of the first Hourglass, the output data of the first Hourglass, and the first-level prediction result. These three channels of data are fused by concat and add, and their scales are different, which reflects the currently popular idea of skip-level (jump) structure.

In the Stacked hourglass network both high-resolution to low-resolution processing and low-resolution to high-resolution processing are symmetrical. The stacked hourglass was tested on MPII and FLIC dataset benchmarks using PCK@0.2 and PCKh@0.5 evaluation metrics in which surpassed all previous SOTA performance. In addition, this work has improved accuracy from 4-5%. on the joints difficult to detect (knees and ankles).

E. DeeperCut: A DEEPER, STRONGER, AND FASTER MULTI-PERSON POSE ESTIMATION MODEL

DeeperCut [33] is a more similar and an upgrade version of the approach presented in DeepCut [52]. DeeperCut has implied a strong body part detectors to generate effective bottom-up proposals for body joints and adapted the extremely deep Residual Network (ResNet [38]) for human body detection whereas DeepCut adapted Fast R-CNN [48] for the task. The proposed keypoints are assembled into a variable number of consistent body part configurations using image-conditioned pairwise terms. Unlike DeepCut, DeeperCut used an incremental optimization strategy that explores the search space more efficiently which leads to both better performance and speed-up factors. Adapting ResNet allowed this work to tackle the problem of vanishing gradients because ResNet tackles the problem bypassing the state though identity layers and modeling residual functions.

Similar to DeepCut, DeeperCut jointly estimates the poses of every individual appeared in an image by minimizing a joint objective based on Integer Linear Programming (ILP). The authors started by making a set of body joint candidates (D) generated by body part detectors and a set of body joint classes(C) such as head, shoulder, and knee in which each candidate's joint has a unary score for every joint class. Adapting ResNet to the fully convolutional model for the sliding window-based body part detection usually brings a stride of 32px which is too coarse for effective joint localization. The authors showed by reducing the stride from 32px to 8px. Besides, to tackling the problem of vanishing gradients in adapting ResNet, DeeeperCut also achieves a large receptive field size which allows them to incorporate context when predicting locations of individual body joints.

After detecting the keypoints, DeeperCut implemented image-conditioned pairwise terms on proposed keypoints. First, an individual in the image is randomly selected and then the location is fixed for each keypoint at its ground truth location using the learned regression. Second, an individual pairwise score-maps will be done and this gets the shape of a cone which extends to the direction of the correct location, but these are visually fuzzy. Finally, by applying an incremental optimization strategy that uses a branch-and-cut algorithm to incrementally solve several pairwise instances to have a valid human pose configuration.

DeeperCut employed the model on LSP, MPII, and COCO dataset using an evaluation metric of AP and mAP which outperformed most of the previous SOTA models except CPM [29] which got similar performance in some cases. Evaluation is done on both single and multi-person pose estimation.

DeeperCut has introduced novel image-conditioned pairwise terms but still needs several minutes per given image (around 4 min/image). However, the pairwise representations

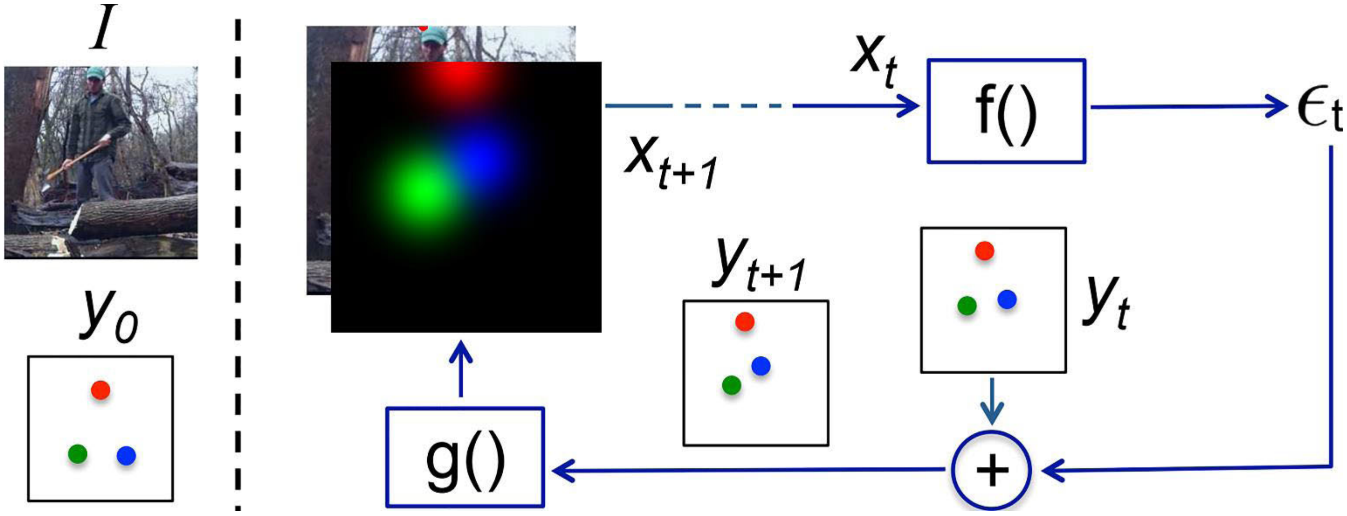

FIGURE 13. Implementation of IEF.

are very hard to regress precisely. Additionally, it implemented the model with a batch size of 1 which increases the instability of the model.

F. IEF: HUMAN POSE ESTIMATION WITH ITERATIVE ERROR FEEDBACK

IEF human pose estimation [72] basically motivated on the concept of prediction, identify what is wrong on this prediction, and correct them iteratively, which is done by a top-down feedback mechanism. IEF employed a framework that extends the hierarchical feature extractor (ConvNet) to include both input and output spaces. In IEF, Error predictions are fed to the initial solution repeatedly and progressively by a self-correcting model as a replacement of directly identifying the keypoints in one go. This framework is called Iterative Error Feedback (IEF) and Fig.13 shows the implementation of IEF for human pose estimation.

On the left side of Fig.13, there is an input composed of the image I and the initially guessed keypoints y_0 (representation of the previous output y_{t-1}). Assume three keypoints to the head (red), the right wrist (green), and the left wrist (blue). Then, define input $X_t = I \oplus g\left(y_{t-1}\right)$, where I represents the image and y_{t-1} is the previous output. The function $f\left(X_t\right)$, modeled as a ConvNet, produces the correction ε_t as output and this output is added to the current output y_t to produce y_{t+1} which means the correction is considered. The function $g\left(y_{t+1}\right)$ converts every keypoint position into one Gaussian heatmap channel so that it can be part of the input with the image for the next iteration. This procedure is done repeatedly and progressively T times until getting a refined y_{t+1} which is very close to the ground truth.

IEF human pose estimation evaluated their performance on two datasets (LSP and MPII) using a single evaluation metric PCKh@0.5. IEF introduced novelty and good work. The functions used, both f and g, are learnable and also, f is a ConvNet. This means f has the ability to learn features over the joint input-output space.

G. REALTIME MULTI-PERSON2D POSE ESTIMATION USING PART AFFINITY FIELDS

Realtime multi-person 2D pose estimation [34] proposed a novelty approach to connect human body parts using Part Affinity Fields (PAF), a non-parametric method, to achieve bottom-up multi-person pose estimation model. The main motivation of this research is identifying the difficulties faced on detecting individual body joints involving multi-person such as the number of people in the image (infinity), the interaction between these people, irregular scale for each individual, increasing complexity, and others.

The overall pipeline and architecture of this model are shown in Fig.14. For a given input image (Fig.14. a), the location of each joint is determined by part confidence maps (Fig.14. b), and the location and orientation of the body parts are determined by PAF (Fig.14. c) a 2D vector that represents the degree of association between the body parts. These body part candidates are associated with the parsing step to perform a set of bipartite matching as shown in Fig.14 (d) and finally, assembled full-body pose because of parsing results in (e).

The two-branch multi-stage CNN network shown in Fig.14 receives an input of a feature map F of an image initialized by the first 10 layers of VGG architecture. The feed-forward model simultaneously predicts confidence maps S (shown in beige) for predicting the location of joints with J confidence maps for each joint ($\mathrm{S} = \mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_\mathrm{J}$) and affinity fields L or a set of 2D vector fields (shown in blue) for encoding parts/limbs association which has C vectors corresponding to each limb ($\mathrm{L} = \mathrm{L}_1, \mathrm{L}_2, \ldots, \mathrm{L}_\mathrm{C}$).

Thus, at the end of the first stage, the network outputs a set of detection confidence maps and part affinity fields. For the consecutive stages, the inputs will be the combination of the two previous stage outputs and the feature map F. Both the confidence maps and the part affinity fields are passed by the greedy inference to have the 2D keypoints for every individual in the image, called Bipartite matching. Furthermore, this work implemented intermediate supervision after each stage

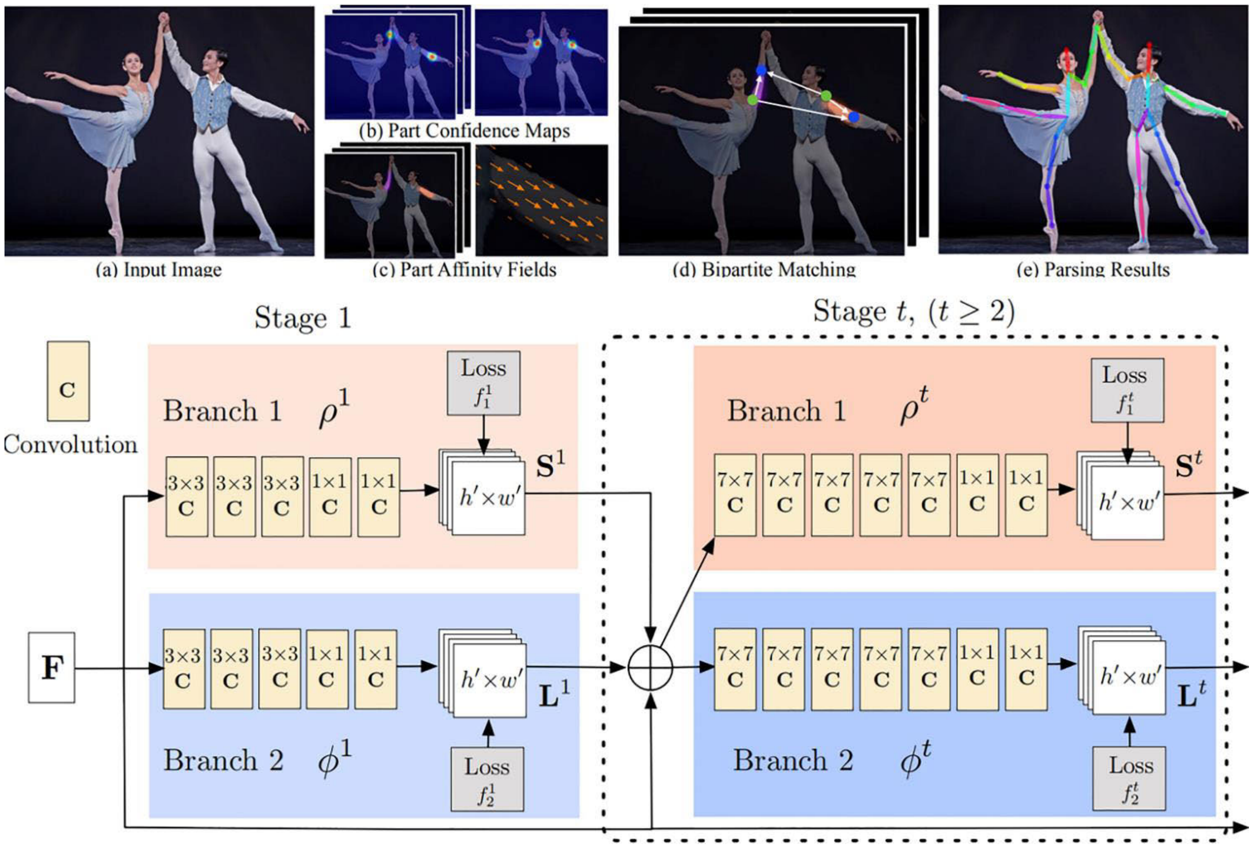

FIGURE 14. Overall pipeline (a - e) and architecture of the two-branch multi-stage CNN.

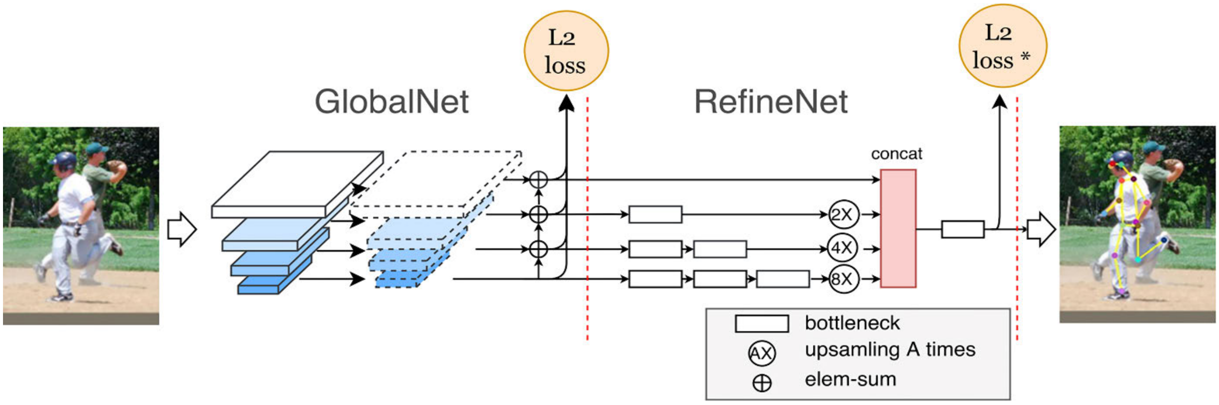

FIGURE 15. The network structure of Cascaded Pyramid Network.

to solve the vanishing gradient's problems by restoring the gradients periodically.

This work is evaluated on COCO and MPII dataset using AP, mAP, and PCKh@0.5 evaluation metrics to achieve the best results compared to the existing SOTA models in terms of performance and efficiency.

H. CPN: CASCADED PYRAMID NETWORK FOR MULTI-PERSON POSE ESTIMATION

Cascaded Pyramid Network (CPN) for multi-person pose estimation model [32] is motivated with the concept of facing the challenging problems which are called "hard keypoints". These include occlusion of keypoints (Occluded by clothes or another person), invisible keypoints, complex backgrounds, etc. The authors proposed a top-down model for multi-person pose estimation with CPN network structure as shown in Fig.15. This CPN network structure is composed of two stages: GlobalNet and RefineNet. Relatively easy keypoints are estimated by the GlobalNet while the hard keypoints estimation is done by RefineNet using online hard keypoint mining loss.

CPN network structure uses a CNN model to identify some human keypoints called Visible easy keypoints, which are relatively easy to detect; for instance Nose, Left elbow, and right hand in the image below. Visible easy keypoints have somewhat a fixed shape and this helps in obtaining texture information which makes it easy to get contextual information around the location of the joints. Then there are visible hard keypoints that are obscured by clothes such as the left knee, right knee, and left hip. Additionally, some joints are hidden and hard to distinguish, not only obscured by clothes, such as the right shoulder in the image shown below. For such hard keypoints, which have no contextual information, increasing the local receptive field is required such that the context information can be further refined. Based on this concept CPN roughly categorized the human body joints into simple parts and difficult parts.

GlobalNet, composed of a forward CNN, is a simple regression model that focuses on easy to detect human keypoints usually eyes, arms, and other easy to detect parts. The purpose of RefineNet is to detect difficult-to-recognize human keypoints, called hard keypoints. RefineNet integrates multiple receptive fields information with the feature maps of the pyramid model generated by GlobalNet. Then finally all feature maps with the same size are concatenated such that a correction for ambiguous keypoints is obtained. RefineNet applies two things to mine the difficult keypoints 1) concat when using features of multiple layers and 2) online hard keypoints mining technology for the second-level network. In general, RefineNet combines low-level features and high-level features through convolution operations.

By the use of RefineNet plus online hard keypoints mining, the model outperformed the previous SOTA models when implementing the model on the COCO dataset using AP and OKS evaluation metrics. CPN exhibits similar properties as stacked hourglass being symmetrical in both processing of high-to-low resolution and low-to-high resolution. It is easy to observe processing from high-resolution to low-resolution as part of a classification network and that it is heavy. Nevertheless, other-way processing (low-resolution to high Resolution) is relatively light.

I. SIMPLE BASELINES FOR HUMAN POSE ESTIMATION AND TRACKING

The main motivation behind simple baselines for human pose estimation and tracking [31] is that most of the recent models on human pose estimation are very complex and look different in structure but achieving very close results. simple baselines proposed a relatively simplified and intuitive model that consists of a few deconvolutional layers at the end of ResNet to estimate the keypoints heatmap. While most human pose estimation models like stacked hourglass [27] and CPN [32] use the structure composed of upsampling and convolution to increase the low-resolution feature map, simple baselines inserts several layers of deconvolution in ResNet which is a very simple way to expand the feature map to the size of the original image to generate the keypoints heatmap as shown in Fig.16.

In this article, both pose estimation and pose tracking are discussed, but our discussion focused on the former.

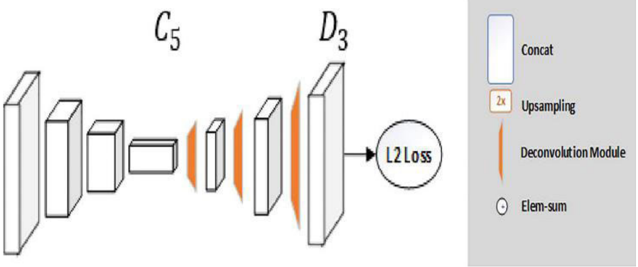

FIGURE 16. Simple Baselines network structure.

As mentioned earlier this model's network structure is straightforward: add several layers of deconvolution after ResNet to generate a heatmap for the individual keypoints. The takeaway from this work is that the more the deconvolution layers, the greater the resolution of the generated heatmap.

Simple baselines achieved better performance compared to the previous works with the COCO dataset using AP evaluation metrics simply and easily. Similar to CPN, high-resolution to low-resolution processing is viewed as part of a classification network (such as ResNet and VGGNet), and this is heavy while processing low-resolution to high-resolution is comparatively light.

J. HRNet: DEEP HIGH-RESOLUTION REPRESENTATION LEARNING FOR HUMAN POSE ESTIMATION

The usual trend applied in human pose estimation is down-sampling high-resolution feature maps to low-resolution and then trying to recover a high-resolution value from low-resolution feature maps. Based on this motivation, this research proposed an intuitive and different model called High-Resolution Net (HRNet) to maintain a high-resolution representation throughout the process [35]. In Stacked hourglass [27] both high-to-low resolution and low-to-high resolution processes are symmetrical. Processing from high-resolution to low-resolution in both CPN [32] and simple baselines [31] considered as part of a classification network by the backbone architecture which is heavy, but the reverse process is relatively light.

There is a high-resolution sub-network at the first stage of this network architecture, as shown in Fig.17. Then gradually a high-to-low resolution sub-networks are added one by one to acquire the output of multiple stages. Finally, the output of multiple resolution sub-networks in parallel are connected. It performs repeated multi-scale fusions such that each high-resolution to low-resolution feature map representation can receive information from other parallel representation branches, again and again, to obtain a more informative high-resolution representation. In the end, the keypoints heatmap of the network output and the spatial resolution are more accurate. Because of repeated multi-scale fusions, HRNet does not need to use intermediate heatmap supervision, unlike the previous works.

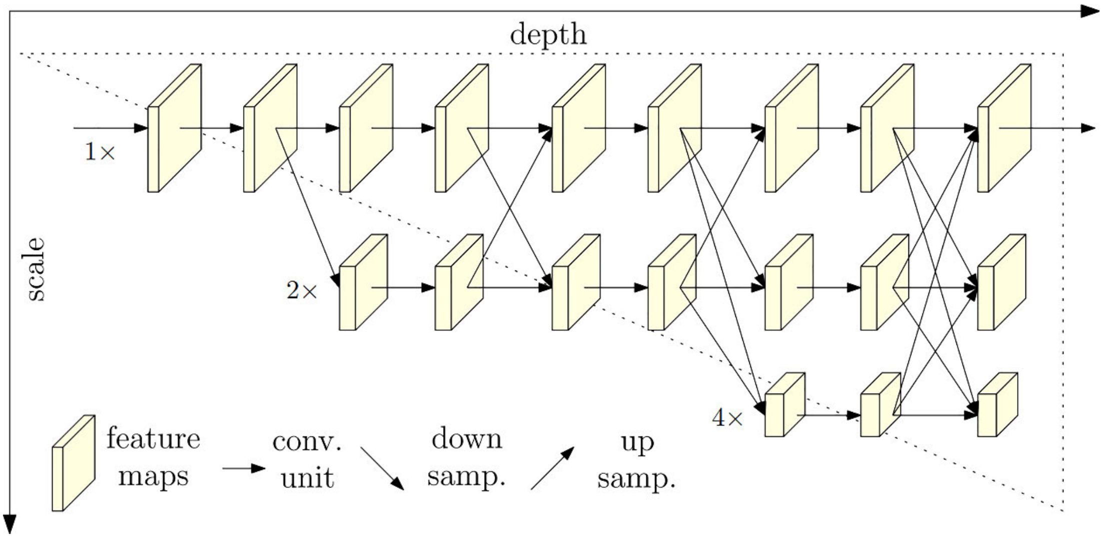

FIGURE 17. The network architecture of HRNet.

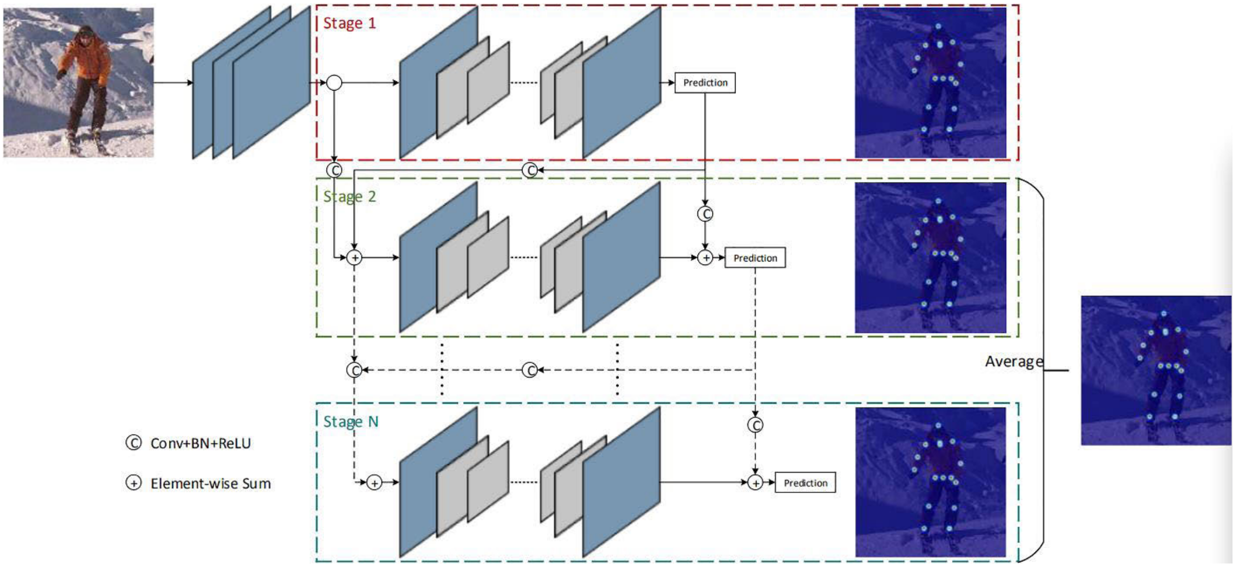

FIGURE 18. CFA network architecture with several stages.

HRNet consists of parallel high-to-low resolution sub-networks with repeated information exchange across multi-resolution sub-networks (multi-scale fusion). The horizontal and vertical directions correspond to the depth of the network and the scale of the feature maps, respectively. There are three scale branches in total. The resolution of the feature map will not change during the forward propagation of each scale branch. Even though there will be information exchange between each scale branch, the three branches are different. For instance, in the forward process, branch 1 (the top branch in the figure) will downsample its feature map and then transfer it to branch 2. Branch 2 will also send the enlarged feature to branch 1 through upsampling. Two operations can be performed in the same stage.

HRNet is evaluated on COCO and MPII dataset using AP, mAP, PCKh@0.5 evaluation metrics to achieve better performance. HRNet introduced the connection of the outputs of high-to-low resolution sub-networks in parallel rather than the usual serial connection. This means it does not require to restore the resolution because high-resolution representations are maintained always.

K. CFA: CASCADE FEATURE AGGREGATION FOR HUMAN POSE ESTIMATION

CFA proposed a cascaded multiple hourglass and aggregates low, medium, and high-level features to better capture local detailed information and global semantic information [26]. The motivation behind CFA network architecture is combining the concept implied in the network architecture of Stacked hourglass [27], CPN [32], and HRNet [35].

The overall network structure of CFA is displayed in Fig.18. CFA consists of multiple hourglass networks that

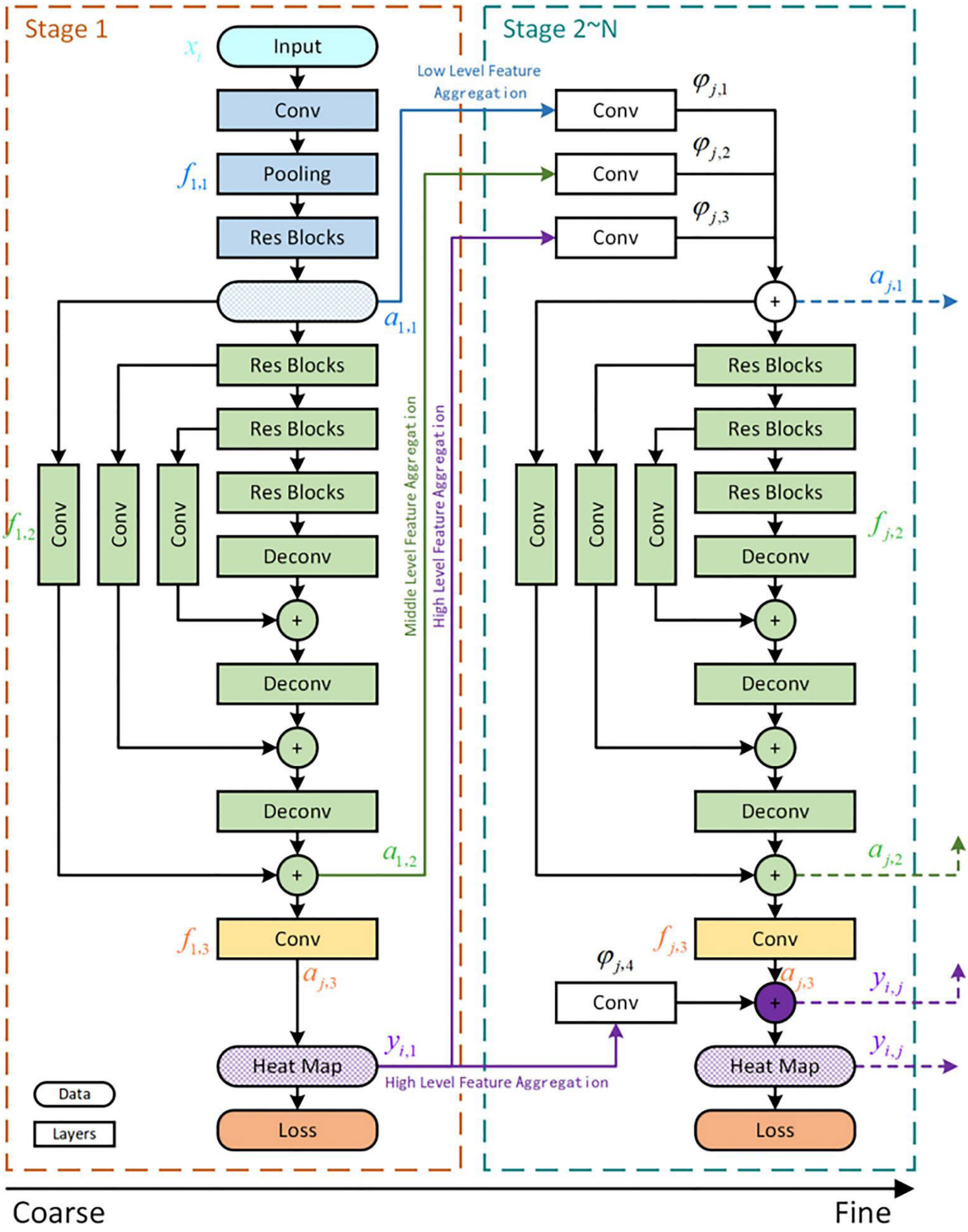

FIGURE 19. Different stages of feature aggregation in CFA.

are summed up by elements. Feature aggregation shows that the hourglass network at each stage will predict the feature map, and the output of the previous layer is used as an input to the next stage at the same time.

In each stage of CFA, ResNet based hourglass network is applied, which is an encoder-decoder model designed based on the hourglass. The basic structure used for the encoder part is ResNet and the connection employed from the encoder to the decoder is highway connection. There are three different feature aggregations in all stages of the CFA model: Low-level feature aggregation, middle-level feature aggregation, and high-level feature aggregation.

Fig.19 briefly describes the feature aggregation between different stages of CFA. Detailed local information is accommodated in low-level features which help them in localizing the exact location of human joints. On the other side, to refine the localization in case of complex backgrounds and partial occlusions, there are high-level features that contain semantic information. Finally, all these different feature aggregations are forwarded as input for the next stage which brings prediction more stable.

CFA evaluated their model on the LIP and MPII datasets using only PCKh@0.5 evaluation metrics. This paper is currently at the top of the 2019 CVPR article based on the MPII dataset PCKh@0.5 evaluation index in the field of single person pose estimation.

L. OccNet: HUMAN POSE ESTIMATION FOR REAL-WORLD CROWDED SCENARIOS

This model proposed in the motivation of estimating the pose of individuals in real-world crowded areas [73]. The challenges of estimating poses in such densely populated areas include people close to each other, mutual occlusions, and partial visibility. The method is a two-stage, top-down approach that localizes the individual first and then performs a single-person pose estimation for every detected person. This model proposed two occlusion detection networks Occlusion Net (OccNet) and Occlusion Net Cross Branch (OccNetCB) as shown in Fig.20, the backbone network is ResNet shown in beige.

In OccNet, to learn a joint representation in the previous layers the network splits after two transposed convolutions. OccNet produces two sets of heatmaps for the location of keypoints per pose: a heatmap for visible keypoints and a heatmap for occluded keypoints. The other architecture, OccNetCB, splits after only one transposed convolution. In OccNetCB, both branches have the opportunity to get information extracted by one another because in OccNetCB the output from both layers is shared.

The model has been evaluated on two datasets annotated on the crowded real-world situation: CrowdPose and JTA datasets using OKS and AP evaluation metrics.

M. DarkPose: DISTRIBUTION-AWARE COORDINATE REPRESENTATION FOR HUMAN POSE ESTIMATION

The main motivation behind the Distribution-Aware Coordinate Representation of Keypoint (DarkPose) is that the coordinate representation of the heatmap [30]. The assumption is that heatmap is never systematically investigated. Based on this concept, the authors have shown design limitations on the existing standard coordinate decoding method, and propose a principled distribution-aware decoding method. In addition to that, an accurate heatmap distribution for the unbiased model training instead of the usual coordinate encoding process is generated (i.e. transforming ground-truth coordinates to heatmaps).

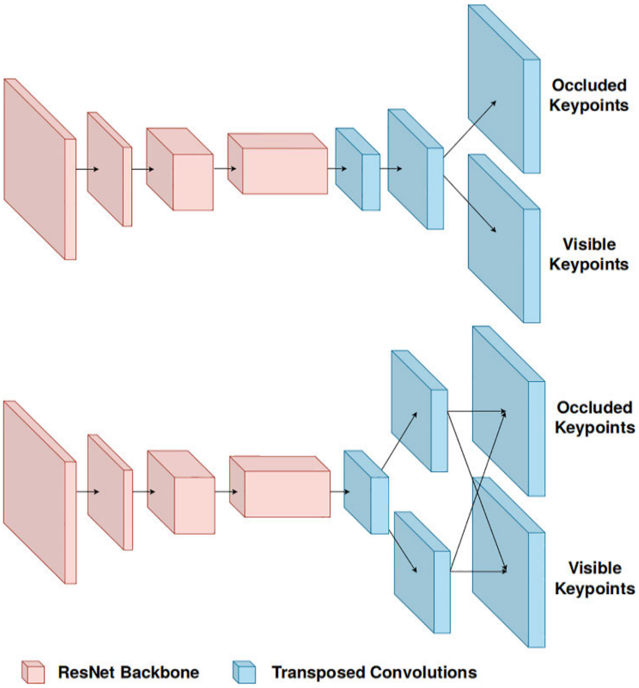

FIGURE 20. The network structure of Occlusion Net (OccNet) and OccNetCB (Cross branch).

Standard label representation in existing methods is coordinate heatmap as a 2-dimensional Gaussian distribution centered at the labeled coordinate of each keypoint of an individual. According to this work, the major obstacle in heatmap label representation and that is quadratic function's computational cost of the input image resolution which restrains CNN based models from processing the typically high-resolution raw imagery data. Hence, there is a need to down-sample all the person bounding box images into a small resolution then fed them to human pose estimation model to predict the location of each keypoint in the original image coordinate space which needs to transform to the original coordinate space, and this brings the problem of sub-pixel localization. Coordinate decoding from heatmap to coordinate is the last prediction of the location with the maximal activation. The network structure of DarkPose is shown in Fig.21.

Coordinate representation, the problem of coordinate encoding and decoding, focused on predicting joint coordinates in a given image. Coordinate decoding is a process of translating a predicted heatmap of each individual's joint into a coordinate in the original image space. Unlike the standard method of considering the second maximum activation to upsample the heatmaps to the original image resolution, DarkPose introduced the heatmap distributional statistics for disclosing the underlying maximum more accurately as shown in Fig.21 and this is employed using Taylor-expansion way. The heatmaps, predicted by a human pose estimation model, usually present multiple peaks around the maximum activation which causes negative effects on the performance

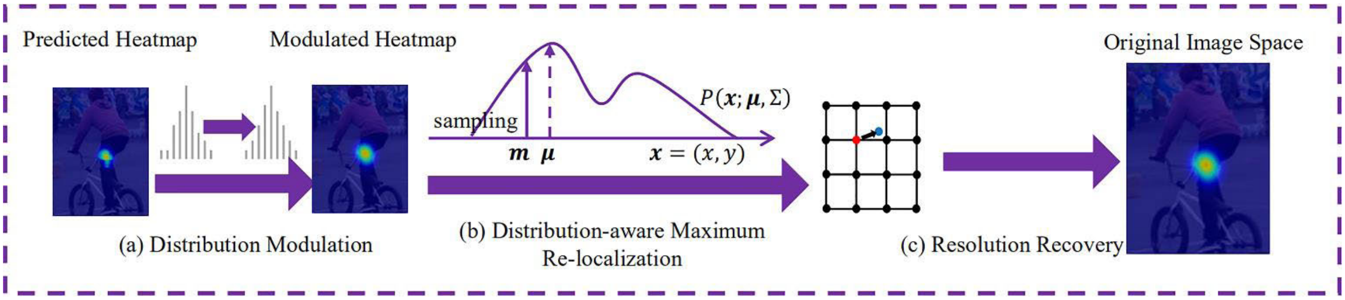
Predicted Heatmap
Modulated Heatmap
sampling
m μ
$P(\boldsymbol{x};\boldsymbol{\mu},\Sigma)$
$\boldsymbol{x}=(x,y)$
Original Image Space
(a) Distribution Modulation
(b) Distribution-aware Maximum Re-localization
(c) Resolution Recovery

FIGURE 21. Overview of the network structure of DarkPose.

TABLE 2. Summary of models.

Models	Backbone architecture	Single / Multi-person	Top-down / Bottom-up	Dataset used	Loss function	Evaluation metrics
DeepPose (2014)	AlexNet	Single Person	Top-down	FLIC, LSP	L2 Loss	PCP, PDJ
ConvNet Pose (2015)	Built-in	Single Person	Top-down	FLIC, MPII	L2 Loss	PCKh@0.5, PCK, PCK@0.05
Convolutional Pose Machines (2016)	VGG structure	Single Person	Top-down	FLIC, LSP, MPII	L2 Loss	PCK@0.1, PCK@0.2, PCKh@0.5
Stacked hourglass (2016)	ResNet	Single Person	Both bottom-up & Top-down	FLIC, MPII	L2 Loss	PCKh@0.5, PCK, PCK@0.2
DeeperCut (2016)	ResNet	Both Single and Multi-Person	Bottom-up	COCO, LSP, MPII	Cross-Entropy loss, L1 Loss	AP, mAP, AUC, PCKh@0.5
Human Pose Estimation with Iterative Error Feedback (2016)	VGG and ConvNet	Single Person	Top-down	LSP, MPII	L2 Loss	PCKh@0.5
Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields (2017)	VGG	Multi-Person	Bottom-up	COCO, MPII	L2 Loss	AP, mAP, PCKh@0.5
Cascaded Pyramid Network for multi-person pose estimation (2018)	ResNet, FPN and Mask R-CNN	Multi-Person	Top-down	COCO	L2 Loss, L2 Loss with online hard keypoints mining	AP, OKS
Simple Baselines for Human Pose Estimation and Tracking (2018)	ResNet	Multi-Person	Top-down	COCO	L2 Loss	AP, mAP
HRNet: Deep High-Resolution Representation Learning for Human Pose Estimation (2019)	ResNet	Multi-Person	Bottom-up	COCO, MPII	L2 Loss	AP, mAP, PCKh@0.5
Cascade Feature Aggregation for Human Pose Estimation (2019)	ResNet	Single Person	Top-down	MPII, LIP	L2 Loss	PCKh@0.5
Human Pose Estimation for Real-World Crowded Scenarios (2019)	ResNet	Multi-Person	Top-down	CrowdPose, JTA	L2 Loss	AP, OKS
Distribution-Aware Coordinate Representation for Human Pose Estimation (2019)	ResNet, HRNet-W32	Single Person	Top-down	COCO, MPII	L2 Loss	PCKh@0.5, PCK, OKS, AP

of the decoding method. To overcome this issue, DarkPose inserted modulating the heatmap distribution before resolution recovery. In coordinate decoding method, three steps employed: heatmap distribution modulation, distribution-aware joint localization by Taylor-expansion as sub-pixel accuracy, and resolution recovery to the original coordinate space. A limitation similar to Coordinate decoding is also observed in coordinate encoding in reducing the resolution. Some of the existing methods start by downsampling given the original image to the model input size. Therefore, in this case, transforming the ground-truth joint coordinates accordingly was necessary before generating heatmaps and this is done by using unbiased sub-pixel centered coordinate encoding.

DarkPose has come up with the concept of problems facing in coordinate representation and the model was evaluated on COCO and MPII using evaluation metrics of PCK and OKS.

V. SUMMARY AND DISCUSSION

This paper reviewed the progress made in pose estimations for human beings with selected and most notable researches made to our knowledge. This discussion started from DeepPose [46], the first well known and has been as reference for most researches in pose estimation progress. Then models have been selected based on their novelty, innovation, the influence made by the model, and other criteria. Table 2 summarizes the models with some criteria.

Human pose estimation deals with the process of inferring poses in an image [4]. To accomplish this objective different kind of techniques have been employed in each model. The techniques can be evaluated with criteria such as the backbone architecture, approaches followed, tracking single or multi-person, the dataset used, loss functions, and evaluation metrics employed.

As shown in Table 2, ResNet [38] nowadays is a default pick as backbone architecture in most models because of its property of solving the problem of vanishing gradients in addition to its great accuracy.

In tracking the number of people in a given image, models are classified as single or multi-person pose estimation. Substantial researches have been carried out in a single person pose estimation with very good results. Even though multi-person pose estimation getting attention, the challenges are still there. These challenges come from the position of each person in the image, visibility of the joints, scale difference, interaction between people, occlusion of joints by clothes, and others.

As shown in Table 2, researchers are preferring Top-down approach instead of bottom-up in most cases. There are also models using both approaches simultaneously.

In datasets selection, COCO and MPII are default picks in recent cases. Especially, COCO is a famous dataset by its property of having very wide human poses and an enormous number of images. LSP and FLIC datasets are also used next to COCO and MPII.

Even though the L_1 loss is not sensitive to outliers, the L_2 loss function is applied in most models to evaluate their learning process. Finally, PCKh@0.5 is the number one evaluation metrics in human pose estimation before mAP and AP.

This article reviewed models focused on determining the full body's pose of individuals. Fascinating researches are also available in discovering only some parts of a human being. For instance: hand pose, head pose, upper body pose, and so on. Additionally, estimating the pose of only children is also presented in [74] research.

VI. CONCLUSION

This paper presented a review of the most outstanding and influential models in human pose estimation progress. As introduced early a 2D human pose estimation has been a fundamental yet challenging problem in computer vision. The main objective of human pose estimation is to localize human anatomical keypoints (e.g., head, shoulder, elbow, wrist, etc.) or joints. This article started by introducing human pose estimation, then classified pose estimation based on tracing the number of people as a single or multi-person. Furthermore, approaches used in pose estimation are explored before discussing its applications and flaws. Finally, some significant papers on pose estimation in both cases of single or multi-person are briefly discussed.

Thus, this article provides a guideline for new readers about human pose estimation. Furthermore, this paper can be a base for research to innovate new models by combining the techniques used in different papers mentioned above. This can be done by changing the backbone architecture or combining the two or three models to create new, or adding new architecture on one of the mentioned papers.

There are very large datasets publicly available on the net. Using these datasets, we have seen substantial progress in 2D human pose estimation with deep learning. However, in addition to the issues discussed in the summary and discussion section, some challenges remain to be addressed in the near future works. Such as i) occlusion of body parts by clothes and other people, ii) interactions between people, iii) human body structure constraints, and iv) barely visible joints are some of the prominent issues that need immense attention to be resolved in the coming works.

REFERENCES

[1] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11969–11978.

[2] S. C. Babu. (2019). *A 2019 Guide to Human Pose Estimation With Deep Learning*. [Online]. Available: https://nanonets.com/blog/human-pose-estimation-2d-guide/

[3] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. NIPS*, 2014, pp. 1736–1744.

[4] D. Mwiti. (2019). *A 2019 Guide to Human Pose Estimation*. [Online]. Available: https://heartbeat.fritz.ai/a-2019-guide-to-human-pose-estimation-c10b79b64b73

[5] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1014–1021, doi: 10.1109/CVPR.2009.5206754.

[6] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 623–630, doi: 10.1109/CVPR.2010.5540156.

[7] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2010, p. 5.

[8] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 588–595.

[9] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. CVPR*, Jun. 2011, pp. 1385–1392, doi: 10.1109/CVPR.2011.5995741.

[10] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013, doi: 10.1109/TPAMI.2012.261.

[11] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 596–603, doi: 10.1109/CVPR.2013.83.

[12] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 723–730, doi: 10.1109/ICCV.2011.6126309.

[13] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (Almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, Sep. 2012, doi: 10.1007/s11263-012-0524-9.

[14] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-H. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," *Sensors*, vol. 16, no. 12, p. 1966, Nov. 2016.

[15] H.-B. Zhang, Q. Lei, B.-N. Zhong, J.-X. Du, and J. Peng, "A survey on human pose estimation," *Intell. Autom. Soft Comput.*, vol. 22, no. 3, pp. 483–489, Jul. 2016, doi: 10.1080/10798587.2015.1095419.

[16] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016, doi: 10.1016/j.neucom.2015.09.116.

[17] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, Dec. 2019, doi: 10.26599/TST.2018.9010100.

[18] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897, doi: 10.1016/j.cviu.2019.102897.

[19] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 52–73, Oct. 2007, doi: 10.1016/j.cviu.2006.10.012.

[20] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009, doi: 10.1109/TPAMI.2008.106.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proc. ECCV*, 2014, pp. 740–755.

[22] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "MPII human pose dataset," in *Proc. CVPR*, Jun. 2014, pp. 3686–3693. Accessed: Apr. 13, 2018. [Online]. Available: http://human-pose.mpi-inf.mpg.de/, doi: 10.1109/CVPR.2014.471.

[23] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[24] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1293–1301.

[25] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 205–214.

[26] Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng, "Cascade feature aggregation for human pose estimation," in *Proc. CVPR*, 2019, pp. 1–18.

[27] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.

[28] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," in *Proc. CVPR*, 2019, pp. 1–10.

[29] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[30] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proc. CVPR*, 2019, pp. 7093–7102.

[31] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 472–487.

[32] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112, doi: 10.1109/CVPR.2018.00742.

[33] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. ECCV*, 2016, pp. 34–50.

[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[35] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[36] K. Simonyan and A. Zisserman, "VGG: Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–14, Apr. 2015.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "AlexNet: ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "ResNet: Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[40] S. Johnson, "LSP: Leeds sports pose dataset," in *Proc. Brit. Mach. Vis. Conf.*, Aug. 2010, pp. 12.1–12.11, doi: 10.5244/C.24.12.

[41] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *Int. J. Comput. Vis.*, vol. 100, no. 1, pp. 16–37, Oct. 2012, doi: 10.1007/s11263-012-0532-9.

[42] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 538–552, Sep. 2012, doi: 10.1109/JSTSP.2012.2196975.

[43] Spagnolo and Paolo, "Proceedings IEEE conference on advanced video and signal based surveillance. AVSS 2003," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Jul. 2003, p. 1, doi: 10.1109/AVSS.2003.1217889.

[44] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010, doi: 10.1109/TPAMI.2009.122.

[45] M. Leo, N. Mosca, P. Spagnolo, P. L. Mazzeo, and A. Distante, "Real-time multi-view event detection in soccer games," in *Proc. 2nd ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Sep. 2008, pp. 1–10, doi: 10.1109/ICDSC.2008.4635729.

[46] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[47] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[48] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[49] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "FPN: Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2017, pp. 936–944.

[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[51] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1347–1355.

[52] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4929–4937.

[53] Z. Zhang, J. Tang, and G. Wu, "Simple and lightweight human pose estimation," in *Proc. CVPR*, 2019, pp. 1–8.

[54] J. Brownlee, "Machine learning mastery: Loss and loss functions for training deep learning neural networks," in *Deep Learning Performance*. Vermont, VIC, Australia: Machine Learning Mastery Pty. Ltd., 2019. [Online]. Available: https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/

[55] R. Parmar, "Common Loss functions in machine learning," in *Towards Data Science*. 2018. [Online]. Available: https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23

[56] Chioka. (2013). *Differences Between L1 and L2 as Loss Function and Regularization*. [Online]. Available: http://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/

[57] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[58] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "SURREAL: Learning from synthetic humans," in *Proc. CVPR*, 2017, pp. 4627–4635. [Online]. Available: https://www.di.ens.fr/willow/research/surreal/data/, doi: 10.1109/CVPR.2017.492.

[59] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10855–10864, doi: 10.1109/CVPR.2019.01112.

[60] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *Proc. ECCV*, 2018, pp. 430–446.

[61] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3674–3681. [Online]. Available: https://bensapp.github.io/flic-dataset.html

[62] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.

[63] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, ''PoseTrack: A benchmark for human pose estimation and tracking,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5167–5176.

[64] J. Hui. (2018). mAP (mean Avarage Precision) for Object Detection. in Medium. [Online]. Available: https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173

[65] S. Chauhan. (2019). Understanding Mean Average Precision for Object Detection (With Python Code). in Medium. [Online]. Available: https://medium.com/analytics-vidhya/map-mean-average-precision-for-object-detection-with-simple-python-demonstration-dcc7b3850a07

[66] P. Cheng and S. He, ''Observer-based finite-time asynchronous control for a class of hidden Markov jumping systems with conic-type nonlinearities,'' *IET Control Theory Appl.*, vol. 14, no. 2, pp. 244–252, Jan. 2020, doi: 10.1049/iet-cta.2019.0443.

[67] P. Cheng, S. He, J. Cheng, X. Luan, and F. Liu, ''Asynchronous output feedback control for a class of conic-type nonlinear hidden Markov jump systems within a finite-time interval,'' *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Mar. 25, 2020, doi: 10.1109/TSMC.2020.2980312.

[68] P. Cheng, J. Wang, S. He, X. Luan, and F. Liu, ''Observer-based asynchronous fault detection for conic-type nonlinear jumping systems and its application to separately excited DC motor,'' *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 3, pp. 951–962, Mar. 2020, doi: 10.1109/TCSI.2019.2949368.

[69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[70] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, ''Efficient object localization using convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.

[71] D. B. West, *Introduction To Graph Theory*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[72] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, ''IEF: Human pose estimation with iterative error feedback,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4733–4742, doi: 10.1109/CVPR.2016.512.

[73] T. Golda, T. Kalb, A. Schumann, and J. Beyerer, ''Human pose estimation for real-world crowded scenarios,'' in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[74] G. Sciortino, G. M. Farinella, S. Battiato, M. Leo, and C. Distante, ''On the estimation of children's poses,'' in *Image Analysis and Processing—ICIAP*, S. Battiato, G. Gallo, R. Schettini, and F. Stanco, Eds. Cham, Switzerland: Springer, 2017, pp. 410–421.

TEWODROS LEGESSE MUNEA was born in Sodo, Ethiopia, in 1985. He received the B.S. degree in computer science from Addis Ababa University, Addis Ababa, Ethiopia, in 2008, and the M.S. degree in computer science and engineering from Ajou University, Suwon, South Korea, in 2015. He is currently pursuing the Ph.D. degree in computer science with the College of Informatics, Xiamen University, Xiamen, China. His current research interests include pose estimation, single and multi-person pose estimation, action recognition, deep learning, and computer vision.

YALEW ZELALEM JEMBRE received the B.Sc. degree in computer science from Addis Ababa University, Ethiopia, in 2007, and the M.Sc. and Ph.D. degrees in computer engineering from Ajou University, in 2012 and 2017, respectively. From 2008 to 2010, he worked as a Fixed Line Next Generation Network (FLNGN) Engineer at Telecom Company ZTE (H.K.) Ethiopian Branch. From September 2017 to December 2018, he was an Assistant Professor with Kyungpook National University, Daegu, South Korea. He is currently working as an Assistant Professor at Keimyung University, Daegu. His research interests include cognitive radio, ad hoc networks, machine learning, and underwater sensor networks. He received the Best Paper Award from the *Journal of Communications and Networks*, in 2015.

HALEFOM TEKLE WELDEGEBRIEL received the B.Sc.Eng. degree in information technology and engineering from the Mekelle Institute of Technology, 2009, and the Master of Technology (M.Tech.) degree in computer and information technology from the College of Engineering, Defence University, Ethiopia, in 2014. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Xiamen University, Xiamen, China. His current research interests include optical character recognition, data mining, and big data using deep learning techniques.

LONGBIAO CHEN received the Ph.D. degrees in computer science from Zhejiang University, China, in 2016, and Sorbonne University, France, in 2018. He worked as a Research Assistant at the Institut Mines-Télécom, France. He is currently an Assistant Professor with the Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, China. His research interests include ubiquitous computing, urban computing, and big data analytics. He has published over 20 articles in top-tier journals and conferences, including ACM UBICOMP, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE TRANSACTIONS ON HUMAN–MACHINE SYSTEMS, and JNCA (Elsevier). He is a Technical Committee Member of ACM SIGSPATIAL China and serves as a PC Member of IEEE UIC Conferences. He received two UBICOMP Honorable Mention Awards, in 2015 and 2016.

CHENXI HUANG is currently an Assistant Professor with Xiamen University. His research interests include image processing, image reconstruction, data fusion, 3-D visualization, machine learning, and so on. He has been an Associate Editor of the *Journal of Medical Imaging and Health Informatics*, since 2019.

CHENHUI YANG received the B.S. and M.S. degrees in automatic control from the National University of Defense Technology, in 1989 and 1992, respectively, and the Ph.D. degree in mechanical engineering from Zhejiang University, in 1995. He started to conduct research on driverless vehicles at the National University of Defense Technology. He has been a Faculty Member with the Computer Science Department, Xiamen University, where he became a Full Professor, in 2005. He was a Visiting Scholar with the Argonne National Laboratory, from 1990 to 2000, and USC, from 2014 to 2015. His research interests include computer vision, graphics, and machine learning, with strong desires to design new products in intelligent transportation, medicine, and industry.